

# Application of Information Theory, Lecture 7

## Relative Entropy

### Handout Mode

Iftach Haitner

Tel Aviv University.

December 1, 2015

# Part I

## **Statistical Distance**

## Statistical distance

- ▶ Let  $p = (p_1, \dots, p_m)$  and  $q = (q_1, \dots, q_m)$  be distributions over  $[m]$
- ▶ Their **statistical distance** (also known as, variation distance) is defined by

$$\text{SD}(p, q) := \frac{1}{2} \sum_{i \in [m]} |p_i - q_i|$$

- ▶ This is simply the  $L_1$  norm between the distribution vectors
- ▶ We will see other “distance” measures for distributions next lecture
- ▶ For  $Z \sim p$  and  $Y \sim q$ , let  $\text{SD}(X, Y) = \text{SD}(p, q)$
- ▶ Claim (HW):  $\text{SD}(p, q) = \max_{S \subseteq [m]} (\sum_{i \in S} p_i - \sum_{i \in S} q_i)$
- ▶ Hence,  $\text{SD}(p, q) = \max_D (\Pr_{X \sim p} [D(X) = 1] - \Pr_{X \sim q} [D(X) = 1])$
- ▶ Interpretation

## Distance from the uniform distribution

- ▶ Let  $X$  be rv over  $[m]$
- ▶  $H(X) \leq \log m$
- ▶  $H(X) = \log m \iff X$  is uniform over  $[m]$

### Theorem 1 (this lecture)

Let  $X$  rv over  $[m]$ . Assume  $H(X) \geq \log m - \varepsilon$ , then

$$\text{SD}(X, \sim [m]) \leq \sqrt{\varepsilon \cdot \frac{\ln 2}{2}} = O(\sqrt{\varepsilon})$$

## Part II

# Relative entropy Distance

# Section 1

## **Definition and Basic Facts**

## Definition

- ▶ For  $p = (p_1, \dots, p_m)$  and  $q = (q_1, \dots, q_m)$ , let

$$D(p||q) = \sum_{i=1}^m p_i \log \frac{p_i}{q_i}$$

$$0 \log \frac{0}{0} = 0, p \log \frac{p}{0} = \infty$$

- ▶ The relative entropy of pair of rv's, is the relative entropy of their distributions.
- ▶ Names: Entropy of  $p$  relative to  $q$ , relative entropy, information divergence, Kullback-Leibler (KL) divergence/distance
- ▶ Many different interpretations
- ▶ Main interpretation: the information we **gained** about  $X$ , if we originally thought  $X \sim q$  and now we learned  $X \sim p$

## Numerical Example

$$D(p\|q) = \sum_{i=1}^m p_i \log \frac{p_i}{q_i}$$

►  $p = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0), q = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$

►  $D(p\|q) = \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{8}} + 0 \log 0 = \frac{1}{4} \cdot (-1) + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 1 = \frac{1}{2}$

►  $D(q\|p) = \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} + \frac{1}{8} \log \frac{\frac{1}{8}}{\frac{1}{4}} + \frac{1}{8} \log \frac{\frac{1}{8}}{0} =$   
 $\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot (-1) + \frac{1}{8} \cdot (-1) + \infty = \infty$



## Supporting the interpretation

- ▶  $X$  rv over  $[m]$
- ▶  $H(X)$  — measure for amount of information we do not have about  $X$
- ▶  $\log m - H(X)$  — measure for information we **do** have about  $X$  (just by knowing its distribution)
- ▶ Example  $X = (X_1, X_2) \sim (\frac{1}{2}, 0, 0, \frac{1}{2})$  over  $\{00, 01, 10, 11\}$
- ▶  $H(X) = 1$ ,  $\log m - H(X) = 2 - 1 = 1$
- ▶ Indeed, we know  $X_1 \oplus X_2$

▶

$$\begin{aligned} H(\sim [m]) - H(p_1, \dots, p_m) &= \log m - H(p_1, \dots, p_m) \\ &= \log m + \sum_i p_i \log p_i = \sum_i p_i (\log p_i - \log \frac{1}{m}) \\ &= \sum_i p_i \log \frac{p_i}{\frac{1}{m}} = D(p \| \sim [m]) \end{aligned}$$

- ▶  $D(X \| \sim [m])$  — **measures** the information we **gained** about  $X$ , if we originally thought it is  $\sim [m]$  and now we learned it is  $\sim p$

## Supporting the interpretation, cont.

- ▶ (generally)  $D(p\|q) \neq H(p) - H(q)$
- ▶  $H(q) - H(p)$  is **not** a good measure for information change
- ▶ Example:  $q = (0.01, 0.99)$  and  $p = (0.99, 0.01)$
- ▶ We were almost sure that  $X = 1$  but learned that  $X$  is almost surely 0
- ▶ But  $H(p) - H(q) = 0$
- ▶ Also,  $H(q) - H(p)$  might be negative
- ▶ We **understand**  $D(p\|q)$  as the information we gained about  $X$ , if we originally thought it is  $\sim q$  and now we learned it is  $\sim p$

# Changing distribution

- ▶ What does it mean: originally thought  $X \sim q$  and now we learned  $X \sim p$ ?

How can a distribution change?

- ▶ Typically, this happens by learning additional information
- ▶  $q_i = \Pr[X = i]$  and  $p_i = \Pr[X = i|E]$
- ▶ Example  $X \sim (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0)$ ; someone saw  $X$  and tells us that  $X \leq 2$
- ▶ The distribution changes to  $X \sim (\frac{2}{3}, \frac{1}{3}, 0, 0)$

- ▶ Another example

$X \backslash Y$	1	2	3	4
0	$\frac{1}{4}$	$\frac{1}{4}$	0	0
1	$\frac{1}{4}$	0	$\frac{1}{4}$	0

- ▶  $Y \sim (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0)$ , but
- ▶  $Y \sim (\frac{1}{2}, \frac{1}{2}, 0, 0)$  conditioned on  $X = 0$
- ▶  $Y \sim (\frac{1}{2}, 0, \frac{1}{2}, 0)$  conditioned on  $X = 1$
- ▶ Generally, a distribution can change if we condition on event  $E$

## Additional properties

- ▶  $0 \log \frac{0}{0} = 0$ ,  $p \log \frac{p}{0} = \infty$  for  $p > 0$
- ▶  $\exists i$  s.t.  $p_i > 0$  and  $q_i = 0$ , then  $D(p\|q) = \infty$
- ▶ If originally  $\Pr[X = i] = 0$ , then it cannot be more than 0 after we learned something.
- ▶ Hence, it make sense to think of it as infinite amount of information learnt
- ▶ Alternatively, we can define  $D(p\|q)$  only for distribution with  $q_i = 0 \implies p_i = 0$   
(recall that  $\Pr[X = i] = 0 \implies \Pr[X = i|E] = 0$ , for any event  $E$ )
- ▶ If  $p_i$  is large and  $q_i$  is small, then  $D(p\|q)$  is large
- ▶  $D(p\|q) \geq 0$ , with equality iff  $p = q$  (hw)

## Example

- ▶  $q = (q_1, \dots, q_m)$  with  $\sum_{i=1}^n q_i = 2^{-k}$  (i.e.,  $n < m$ )
- ▶  $p_i = \begin{cases} q_i/2^{-k}, & 1 \leq i \leq n \\ 0, & \text{otherwise.} \end{cases}$
- ▶  $p = (p_1, \dots, p_m)$  — the distribution of  $q$  conditioned on the event  $i \in [n]$
- ▶  $D(p\|q) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i} = \sum_{i=1}^n p_i \log 2^k = \sum_{i=1}^n p_i k = k$
- ▶ We gained  $k$  bits of information
- ▶ Example:  $\sum_{i=1}^n q_i = \frac{1}{2}$ , and we were told that  $i \leq n$  or  $i > n$ , we got one bit of information

## Section 2

# Axiomatic Derivation

## Axiomatic derivation

Let  $\tilde{D}$  is a continuous and symmetric (wrt each distribution) function such that

1.  $\tilde{D}(p \parallel \sim [m]) = \log m - H(p)$
2.  $\tilde{D}((p_1, \dots, p_m) \parallel (q_1, \dots, q_m)) = \tilde{D}((p_1, \dots, p_{m-1}, \alpha p_m, (1 - \alpha)p_m) \parallel (q_1, \dots, q_{m-1}, \alpha q_m, (1 - \alpha)q_m))$ , for any  $\alpha \in [0, 1]$

then  $\tilde{D} = D$ .

Interpretation

Proof:

- ▶  $\tilde{D}(p \parallel q) = D((\alpha_{1,1}p_1, \dots, \alpha_{1,k_1}p_1, \dots, \alpha_{m,1}p_m, \dots, \alpha_{m,k_m}p_m) \parallel (\alpha_{1,1}q_1, \dots, \alpha_{1,k_1}q_1, \dots, \alpha_{m,1}q_m, \dots, \alpha_{m,k_m}q_m))$ , for  $\sum_j \alpha_{i,j} = 1$  and  $\alpha_{i,j} \geq 0$

- ▶ Taking  $\alpha$ 's s.t.  $\alpha_{i,1} = \alpha_{i,2} \dots, \alpha_{i,k_i} = \alpha_i$  and  $\alpha_i q_i = \frac{1}{M}$ , it follows that

$$\begin{aligned}\tilde{D}(p \parallel q) &= \log M - H((\alpha_{1,1}p_1, \dots, \alpha_{1,k_1}p_1, \dots, \alpha_{m,1}p_m, \dots, \alpha_{m,k_m}p_m)) \\ &= \sum p_i \log M + \sum_i p_i \log \alpha_i p_i = \sum_i p_i (\log M + \log \frac{p_i}{q_i M}) = \sum_i p_i \log \frac{p_i}{q_i}.\end{aligned}$$

- ▶ Zeros and non-rational  $q_i$ 's are dealt by continuity

## Section 3

# Relation to Mutual Information



## Mutual information as expected relative entropy

- ▶ Let  $X \sim (q_1, \dots, q_m)$  over  $[m]$ , and  $Y$  be rv over  $\{0, 1\}$
- ▶  $(X|Y=0) \sim p_0 = (p_{0,1}, \dots, p_{0,m})$ ,  $p_{0,i} = \Pr[X=i|Y=0]$
- ▶  $(X|Y=1) \sim p_1 = (p_{1,1}, \dots, p_{1,m})$ ,  $p_{1,i} = \Pr[X=i|Y=1]$
- ▶ If we learned  $Y=j$ , we gained  $D(p_j||q)$

$$\begin{aligned} \mathbb{E}_Y [D(p_Y||q)] &= \Pr[Y=0] \cdot D(p_{0,1}, \dots, p_{0,m}||q_1, \dots, q_m) \\ &\quad + \Pr[Y=1] \cdot D(p_{1,1}, \dots, p_{1,m}||q_1, \dots, q_m) \\ &= \Pr[Y=0] \cdot \sum_i p_{0,i} \log \frac{p_{0,i}}{q_i} + \Pr[Y=1] \cdot \sum_i p_{1,i} \log \frac{p_{1,i}}{q_i} \\ &= \Pr[Y=0] \cdot \sum_i p_{0,i} \log p_{0,i} + \Pr[Y=1] \cdot \sum_i p_{1,i} \log p_{1,i} \\ &\quad - \Pr[Y=0] \cdot \sum_i p_{0,i} \log q_i - \Pr[Y=1] \cdot \sum_i p_{1,i} \log q_i \\ &= -H(X|Y) - \sum_i (\Pr[Y=0] \cdot p_{0,i} + \Pr[Y=1] \cdot p_{1,i}) \log q_i \\ &= -H(X|Y) + H(X) = I(X; Y) \end{aligned}$$

## Equivalent definition for mutual information

►  $(X, Y) \sim p$ , then  $I(X; Y) = D(p \| p_X p_Y)$

► Interpretation

► Proof:

$$\begin{aligned} D(p \| p_X p_Y) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p_X(x) p_Y(y)} \\ &= \sum_{x,y} p(x, y) \log \frac{p_{X|Y}(x|y)}{p_X(x)} \\ &= - \sum_{x,y} p(x, y) \log p_X(x) + \sum_{x,y} p(x, y) \log p_{X|Y}(x|y) \\ &= H(X) + \sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log p_{X|Y}(x|y) \\ &= H(X) - H(X|Y) = I(X; Y) \end{aligned}$$

► We will later see the relation between the above two facts.

## Section 4

# **Relation to Data Compression**

## Wrong code

### Theorem 2

Let  $p$  and  $q$  be distributions over  $[m]$ , and let  $C$  be code with

$\ell(i) = |C(i)| = \left\lceil \log \frac{1}{q_i} \right\rceil$ . Then

$$H(p) + D(p\|q) \leq \mathbb{E}_{i \leftarrow p} [\ell(i)] \leq H(p) + D(p\|q) + 1$$

- ▶ Recall that  $H(q) \leq \mathbb{E}_{i \leftarrow q} [\ell(i)] \leq H(q) + 1$ .
- ▶ Proof of upperbound (upperbound is proved similarly)

$$\begin{aligned} \mathbb{E}_{i \leftarrow p} [\ell(i)] &= \sum_i p_i \left\lceil \log \frac{1}{q_i} \right\rceil < \sum_i p_i (\log \frac{1}{q_i} + 1) \\ &= 1 + \sum_i p_i (\log \frac{p_i}{q_i} \frac{1}{p_i}) = 1 + \sum_i p_i (\log \frac{p_i}{q_i}) + \sum_i p_i (\log \frac{1}{p_i}) \\ &= 1 + D(p\|q) + H(p) \end{aligned}$$

- ▶ Can there be a (close) to optimal code for  $q$  that is better for  $p$ ? HW

## Section 5

# Conditional Relative Entropy

# Conditional relative entropy

## Definition 3

For two distributions  $p$  and  $q$  over  $\mathcal{X} \times \mathcal{Y}$ :

$$\begin{aligned} D(p_{\mathcal{Y}|\mathcal{X}} \| q_{\mathcal{Y}|\mathcal{X}}) &:= \sum_{x \in \mathcal{X}} p_{\mathcal{X}}(x) \cdot \sum_{y \in \mathcal{Y}} p_{\mathcal{Y}|\mathcal{X}}(y|x) \log \frac{p_{\mathcal{Y}|\mathcal{X}}(y|x)}{q_{\mathcal{Y}|\mathcal{X}}(y|x)} \\ &= \mathbb{E}_{(X,Y) \sim p(X,Y)} \left[ \log \frac{p_{\mathcal{Y}|\mathcal{X}}(Y|X)}{q_{\mathcal{Y}|\mathcal{X}}(Y|X)} \right] \end{aligned}$$

- Let  $(X_p, Y_p) \sim p$  and  $(X_q, Y_q) \sim q$ , then

$$D(p_{\mathcal{Y}|\mathcal{X}} \| q_{\mathcal{Y}|\mathcal{X}}) = \mathbb{E}_{x \leftarrow X_p} [D(Y_q|_{X_p=x} \| Y_q|_{X_q=x})]$$

- Example:  $p =$ 

$X \backslash Y$	0	1
0	$\frac{1}{8}$	$\frac{1}{8}$
1	$\frac{1}{4}$	$\frac{1}{2}$

 $q =$ 

$X \backslash Y$	0	1
0	$\frac{1}{8}$	$\frac{1}{4}$
1	$\frac{1}{2}$	$\frac{1}{8}$

$$\begin{aligned} D(p_{\mathcal{Y}|\mathcal{X}} \| q_{\mathcal{Y}|\mathcal{X}}) &= \frac{1}{4} \cdot D\left(\left(\frac{1}{2}, \frac{1}{2}\right) \parallel \left(\frac{1}{3}, \frac{2}{3}\right)\right) + \frac{3}{4} \cdot D\left(\left(\frac{1}{3}, \frac{2}{3}\right) \parallel \left(\frac{4}{5}, \frac{1}{5}\right)\right) \\ &= \dots \end{aligned}$$

## Chain rule

### Claim 4

For any two distributions  $p$  and  $q$  over  $\mathcal{X} \times \mathcal{Y}$ , it holds that

$$D(p\|q) = D(p_{\mathcal{X}}\|q_{\mathcal{X}}) + D(p_{\mathcal{Y}|\mathcal{X}}\|q_{\mathcal{Y}|\mathcal{X}})$$

► Proof:

$$\begin{aligned} D(p\|q) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p_{\mathcal{X}}(x)p_{\mathcal{Y}|\mathcal{X}}(y|x)}{q_{\mathcal{X}}(x)q_{\mathcal{Y}|\mathcal{X}}(y|x)} \\ &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p_{\mathcal{X}}(x)}{q_{\mathcal{X}}(x)} + \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p_{\mathcal{Y}|\mathcal{X}}(y|x)}{q_{\mathcal{Y}|\mathcal{X}}(y|x)} \\ &= D(p_{\mathcal{X}}\|q_{\mathcal{X}}) + D(p_{\mathcal{Y}|\mathcal{X}}\|q_{\mathcal{Y}|\mathcal{X}}) \end{aligned}$$

► It follows that for  $(X; Y) \sim p$ :  $I(X, Y) = D(p\|p_X p_Y) = D(p_X\|p_X) + \mathbb{E}_{x \leftarrow X} [D(p_{Y|X=x}, p_Y)] = \mathbb{E}_{x \leftarrow X} [D(p_{Y|X=x}, p_Y)]$

## Section 6

# Data-processing inequality



# Data-processing inequality

## Claim 5

For any rv's  $X$  and  $Y$  and function  $f$ , it holds that  $D(f(X)\|f(Y)) \leq D(X\|Y)$ .

- ▶ Analogues to  $H(X) \geq H(f(X))$
- ▶ Proof:
- ▶  $D(X, f(X)\|Y, f(Y)) = D(X\|Y)$
- ▶  $D(X, f(X)\|Y, f(Y)) = D(f(X)\|f(Y)) + \mathbb{E}_{z \leftarrow f(X)} [D(X|_{f(X)=z}\|Y|_{f(X)=z})] \geq D(f(X)\|f(Y))$
- ▶ Hence,  $D(f(X)\|f(Y)) \leq D(X\|Y)$ .

## Section 7

# **Relation to Statistical Distance**

## Relation to statistical distance

- ▶  $D(p\|q)$  is used many time to measure the distance from  $p$  to  $q$
- ▶ It is **not** a distance in the mathematical sense:  $D(p\|q) \neq D(q\|p)$  and no triangle inequality
- ▶ However,

### Theorem 6

$$SD(p, q) \leq \sqrt{\frac{\ln 2}{2} \cdot D(p\|q)}$$

- ▶ Corollary: For rv  $X$  over  $[m]$  with  $H(X) \geq m - \varepsilon$ , it holds that
$$SD(X, \sim [m]) \leq \sqrt{\frac{\ln 2}{2} \cdot (m - H(X))} = \sqrt{\frac{\ln 2}{2} \cdot \varepsilon}$$
- ▶ Other direction is incorrect:  $SD(p, q)$  might be small but  $D(p\|q) = \infty$
- ▶ Does  $SD(p, \sim [m])$  being small imply  $D(p\| \sim [m]) = \log m - H(p)$  is small?

HW

## Proving Thm 6, boolean case

► Let  $p = (\alpha, 1 - \alpha)$  and  $q = (\beta, 1 - \beta)$  and assume  $\alpha \geq \beta$

►  $SD(p, q) = \alpha - \beta$

► We will show that

$$D(p||q) = \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta} \geq \frac{4}{2 \ln 2} (\alpha - \beta)^2 = \frac{2}{\ln 2} SD(p, q)^2$$

► Let  $g(\alpha, \beta) = \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta} - \frac{4}{2 \ln 2} (\alpha - \beta)^2$

$$\begin{aligned} \frac{\partial g(x, y)}{\partial y} &= -\frac{x}{y \ln 2} + \frac{1 - x}{(1 - y) \ln 2} - \frac{4}{2 \ln 2} 2(y - x) \\ &= \frac{y - x}{y(1 - y) \ln 2} - \frac{4}{\ln 2} (y - x) \end{aligned}$$

► Since  $y(1 - y) \leq \frac{1}{4}$ ,  $\frac{\partial g(x, y)}{\partial y} \leq 0$  for  $y < x$ .

► Since  $g(x, x) = 0$ ,  $g(x, y) \geq 0$  for  $y < x$ .  $\square$

## Proving Thm 6, general case

- ▶ Let  $\mathcal{U} = \text{Supp}(p) \cup \text{Supp}(q)$
- ▶ Let  $\mathcal{S} = \{u \in \mathcal{U} : p(u) > q(u)\}$
- ▶  $\text{SD}(p, q) = \Pr_p[\mathcal{S}] - \Pr_q[\mathcal{S}]$  (by homework)
- ▶ Let  $P \sim p$ , and let the indicator  $\hat{P}$  be 1 iff  $P \in \mathcal{S}$ .
- ▶ Let  $Q \sim q$ , and let the indicator  $\hat{Q}$  be 1 iff  $Q \in \mathcal{S}$ .
- ▶  $\text{SD}(\hat{P}, \hat{Q}) = \Pr[P \in \mathcal{S}] - \Pr[Q \in \mathcal{S}] = \text{SD}(p, q)$

$$\begin{aligned} D(p \| q) &\geq D(\hat{P} \| \hat{Q}) && \text{(data-processing inequality)} \\ &\geq \frac{2}{\ln 2} \cdot \text{SD}(\hat{P}, \hat{Q})^2 && \text{(the Boolean case)} \\ &= \frac{2}{\ln 2} \cdot \text{SD}(p, q)^2. \quad \square \end{aligned}$$

## Section 8

# Conditioned Distributions

# Main theorem

## Theorem 7

Let  $X_1, \dots, X_k$  be iid over  $\mathcal{U}$ , and let  $Y = (Y_1, \dots, Y_k)$  be rv over  $\mathcal{U}^k$ . Then  $\sum_{j=1}^k D(Y_j \| X_j) \leq D(Y \| (X_1, \dots, X_k))$ .

For rv  $Z$ , let  $Z(z) = \Pr[Z = z]$ .

We prove for  $k = 2$ , general case follows similar lines. Let  $X = (X_1, X_2)$

$$\begin{aligned} D(Y \| X) &= \sum_{\mathbf{y} \in \mathcal{U}^2} Y(\mathbf{y}) \log \frac{Y(\mathbf{y})}{X(\mathbf{y})} = \sum_{\mathbf{y}=(y_1, y_2)} Y(\mathbf{y}) \log \frac{Y_1(y_1)}{X_1(y_1)} \frac{Y_2(y_2)}{X_2(y_2)} \frac{Y(\mathbf{y})}{Y_1(y_1) Y_2(y_2)} \\ &= \sum_{\mathbf{y}=(y_1, y_2)} Y(\mathbf{y}) \log \frac{Y_1(y_1)}{X_1(y_1)} + \sum_{\mathbf{y}=(y_1, y_2)} Y(\mathbf{y}) \log \frac{Y_2(y_2)}{X_2(y_2)} \\ &\quad + \sum_{\mathbf{y}=(y_1, y_2)} Y(\mathbf{y}) \log \frac{Y(\mathbf{y})}{Y_1(y_1) Y_2(y_2)} \\ &= D(Y_1 \| X_1) + D(Y_2 \| X_2) + I(Y_1; Y_2) \geq D(Y_1 \| X_1) + D(Y_2 \| X_2) \end{aligned}$$

# Conditioning distributions, relative entropy case

## Theorem 8

Let  $X_1, \dots, X_k$  be iid over  $\mathcal{X}$ , let  $X = (X_1, \dots, X_k)$  and let  $W$  be an event (i.e., Boolean rv). Then  $\sum_{j=1}^k D((X_j|_W) \| X_j) \leq D((X|_W) \| X) \leq \log \frac{1}{\Pr[W]}$ .

$$\sum_{j=1}^k D((X_j|_W) \| X_j) \leq D((X|_W) \| X) \quad (\text{Thm 7})$$

$$\begin{aligned} &= \sum_{\mathbf{x} \in \mathcal{X}^k} (X|_W)(\mathbf{x}) \log \frac{(X|_W)(\mathbf{x})}{X(\mathbf{x})} \\ &= \sum_{\mathbf{x} \in \mathcal{X}^k} (X|_W)(\mathbf{x}) \log \frac{\Pr[W|X = \mathbf{x}]}{\Pr[W]} \quad (\text{Bayes}) \\ &= \log \frac{1}{\Pr[W]} + \sum_{\mathbf{x} \in \mathcal{X}^k} (X|_W)(\mathbf{x}) \log \Pr[W|X = \mathbf{x}] \\ &\leq \log \frac{1}{\Pr[W]} \end{aligned}$$



# Conditioning distributions, statistical distance case

## Theorem 9

Let  $X_1, \dots, X_k$  be iid over  $\mathcal{X}$  and let  $W$  be an event. Then

$$\sum_{j=1}^k \text{SD}((X_j|_W), X_j)^2 \leq \log \frac{1}{\Pr[W]}.$$

Proof: follows by Thm 6, and Thm 7.  $\square$

Using  $(\sum_{j=1}^k a_j)^2 \leq k \cdot \sum_{j=1}^k a_j^2$ , it follows that

## Corollary 10

$\sum_{j=1}^k \text{SD}((X_j|_W), X_j) \leq \sqrt{k \log(\frac{1}{\Pr[W]})}$ , and

$\mathbb{E}_{j \leftarrow k} \text{SD}((X_j|_W), X_j) \leq \sqrt{\frac{1}{k} \log(\frac{1}{\Pr[W]})}$

Extraction

## Numerical example

- ▶ Let  $X = (X_1, \dots, X_k) \leftarrow \{0, 1\}^{40}$  and let  $f: \{0, 1\}^{40} \mapsto 0$  be such that  $\Pr[f(X) = 0] = 2^{-10}$ .
- ▶  $E_{j \leftarrow [40]} \text{SD}((X_j|_{f(X)=0}), \sim \{0, 1\}) \leq \sqrt{\frac{1}{40} \cdot 10} = \frac{1}{2}$
- ▶ Typical bits are not too biased, even when conditioning on a very unlikely event.

## Extension

### Theorem 11

Let  $X = (X_1, \dots, X_k)$ ,  $T$  and  $V$  be rv's over  $\mathcal{X}^k$ ,  $\mathcal{T}$  and  $\mathcal{V}$  respectively. Let  $W$  be an event and assume that the  $X_i$ 's are iid conditioned on  $T$ . Then

$$\sum_{j=1}^k D((TVX_j)|_w || (TV)|_w X'_j(T)) \leq \log \frac{1}{\Pr[W]} + \log |\text{Supp}(V|_w)|,$$

where  $X'_j(t)$  is distributed according to  $X_j|_{T=t}$ .

Interpretation.

## Proving Thm 11

Let  $X = (X_1, \dots, X_k)$ ,  $T$  and  $V$  be rv's over  $\mathcal{X}^k$ ,  $\mathcal{T}$  and  $\mathcal{V}$  respectively, such that  $X_i$ 's are iid conditioned on  $T$ . Let  $W$  be an event and let  $X_j'(t)$  be distributed according to the distribution of  $X_j|_{T=t}$ .

$$\begin{aligned} & \sum_{j=1}^k D((TVX_j)|_W || (TV)|_W X_j'(T)) \\ &= \mathbb{E}_{(t,v) \leftarrow (TV)|_W} \left[ \sum_{j=1}^k D(X_j|_{W, V=v, T=t} || (X_j|_{T=t})) \right] \end{aligned} \quad \text{(chain rule)}$$

$$\begin{aligned} &= \mathbb{E}_{(t,v) \leftarrow (TV)|_W} \left[ \sum_{j=1}^k D(\underbrace{(X_j|_{W, V=v})}_{W'}|_{T=t} || (X_j|_{T=t})) \right] \\ &\leq \mathbb{E}_{(t,v) \leftarrow (TV)|_W} \left[ \log \frac{1}{\Pr[W \wedge V=v | T=t]} \right] \end{aligned} \quad \text{(Thm 8)}$$

$$\leq \log \mathbb{E}_{(t,v) \leftarrow (TV)|_W} \frac{1}{\Pr[W \wedge V=v | T=t]} \quad \text{(Jensen's inequality)}$$

$$= \log \sum_{(t,v) \in \text{Supp}((TV)|_W)} \frac{\Pr[T=t]}{\Pr[W]} \leq \log \frac{|\text{Supp}(V|_W)|}{\Pr[W]}.$$

□